The Study and Analysis of Classification Algorithm for Animal Kingdom Dataset

E. Bhuvaneswari*[1], V. R. Sarma Dhulipala[2]

Assistant Professor, Department of CSE, E.S Engineering College, Villupuram, Tamil Nadu, India, Assistant Professor, Department of Physics , Anna University, BIT Campus, Tiruchirappalli-620024, India

ebhuvaneswari@gmail.com, [2] dvrsarma@gmail.com

Abstract

The study of evolution in the animal world is immensely diverse. Evolution of animals can be categorized using data mining tools such as Weka. It is one of the freely available tools which provide a single platform to combine classification, clustering, association, validation and visualization. Classification is the arrangement of objects, ideas, or information into groups, the members of which have one or more characteristics in common. Classification makes things easier to find, identify, and study. Taking diversity into account the number of species is classified using the attributes in weka. The animal kingdom is categorized as vertebrates and invertebrates. In this paper animal kingdom data set is developed by collecting data from A to Z vertebrate's animal kingdom repository. Data set consists of 51 instances with 6 attributes. The considered attributes are name, weight, size, lifespan, origin, and group. The dataset is trained and tested using remove percentage filter. Partitioned data set are evaluated individually using weka algorithms and the results are compared using error rate and accuracy rate. The results are compared and verified using Knowledge flow environment.

Keywords

Machine Learning; Data Mining; WEKA; Classification; Knowledge Flow Experimenter; Animal Kingdom Data Set

Introduction

There is a staggering increase in the population and evolution of living things in the environment. Populations are groups of individuals belonging to the same region. Populations, like individual organisms, have unique attributes such as: growth rate, age structure, sex ratio, mortality rate [2]. The first and largest category in population evolution is the Kingdom. There are five kingdoms in our environment. There are over 1 million different species of animals that have been identified and classified and perhaps millions and more than that have not been classified. It is mainly categorized into two forms vertebrates and invertebrates. Vertebrates, the animals in higher order compared with invertebrates. Vertebrates are divided into five different groups: mammals, birds, amphibians, reptiles and fish. We classify living things according to the characteristics they share [1]. To study different types of animals, it is convenient, classify them by common characteristics. The main focus of this paper is to classify the animal based on the attributes [18]. Weka is one of the frameworks for classification that contains many well-known data mining algorithms. Classification in weka is made by considering the attributes such as origin life span, weight, size, color etc., Although each of these groups of animals has unique characteristics, they have some common characteristics as well [2].

Weka is a machine learning tool which complements data mining. An understanding of algorithms is combined with detailed knowledge of the datasets. Data sets in weka are validation, training and test set. The data sets to weka are in three forms 1. Direct ataset. 2. Pre categorized dataset 3. Raw data set. In this paper pre categorized datasets are provided to weka to analyze the performance of algorithms. The performance of classification is analyzed using classified instances, error rate, and kappa statistics.

It is widely known that classifiers possess different performance measures. Each classifier may unknowingly work better in training and testing set. The performances of the data sets are tested using different algorithms.

Data Mining Tool: WEKA

Data Mining is the process of extracting information from large data sets through different techniques [3]. Data Mining, popularly called as knowledge discovery

in large data by analyzing and accessing statistical and from data base. In this paper we have used WEKA, a Data Mining tool for classification techniques. Weka provides the required data mining functions and methodologies. The data format for WEKA is MS Excel and ARFF formats respectively. Weka a machine learning workbench implements algorithms for data preprocessing, classification, regression, clustering and association rules [4]. Implementation in weka is classified as:

1. Implementation scheme for classification;
2. Implementation schemes for numeric prediction;
3. Implemented meta-schemes.

Learning methods in weka are called classifiers which contain tunable parameters that can be accessed through a property sheet or object editor. The exploration modes in weka allow data preprocessing, learning, data processing, and attribute selection and data visualization modules in an environment that encourages initial exploration of data. Data are pre processed using Remove useless filter. It removes the largely varying, less varying data in the data sets [8]. Remove percentage filter is used for training and testing the data set.

Data Set

Records of data base have been created in Excel data sheet and saved in the format of CSV (Comma Separated Value format) that converted to the WEKA accepted of ARFF by using command line premier of WEKA. Predominant vertebrate animal data sets are taken for classification. The records of data base consist of 6 attributes, from which 5 attributes were selected using remove useless filter which filters the unwanted attributes in the data set [23]. Only 60% of the overall data is used as a training set and the remaining is used as test set [7].

Training Set And Test Set

Full data set is trained using remove percentage filter in the pre- process panel. Full data set is again loaded for testing the data set.

Testing set is prepared using invert selection property to true values by applying the correct percentage filter. Remove Useless filter: it removes the large and less varying data in the entire data set. The considered attributes are name, average weight, average size, life span, origin [6]. Remove percentage filter is used to split the overall data set into training and tested data set. In our data set, name is the largely varying attribute. Remove useless filter to remove the name attribute in the data set.

Classification Methods

NAÏVE BAYES:

In Naïve Bayes classifier attributes are conditionally independent [10]. This greatly reduces the computation cost. It counts only the class distribution.

There are m classes C_1, C_2… C_m. With tuples $\mathbf{X} = (x_1, x_2 \ldots x_n)$, The Classification of such classes is derived using the maximum posteriori, i.e., the maximal $P(C_i|\mathbf{X})$. This can be derived from Baye's theorem [16]. P(X) delete constant for all classes, only needs to be maximized. The goal of this classification is to correctly predict the value of a designated discrete class variable given a vector of attribute using 10 fold cross validation [24]. Naïve Bayes classifier is applied to trained and test set and the performance is evaluated individually with kappa statistics, error rate.

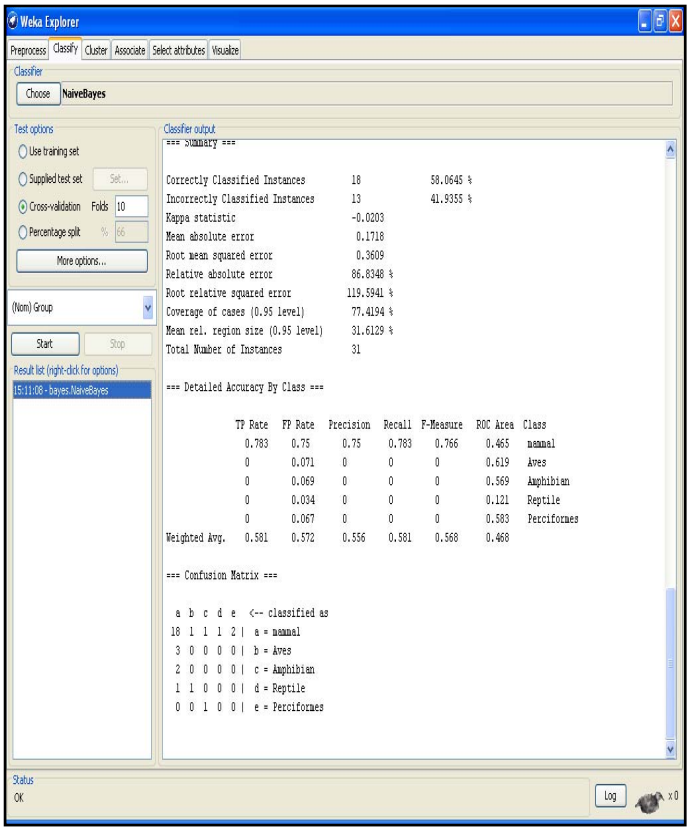

FIG. 1 SIMULATION RESULT FOR TRAINING SET: NAÏVE BAYES

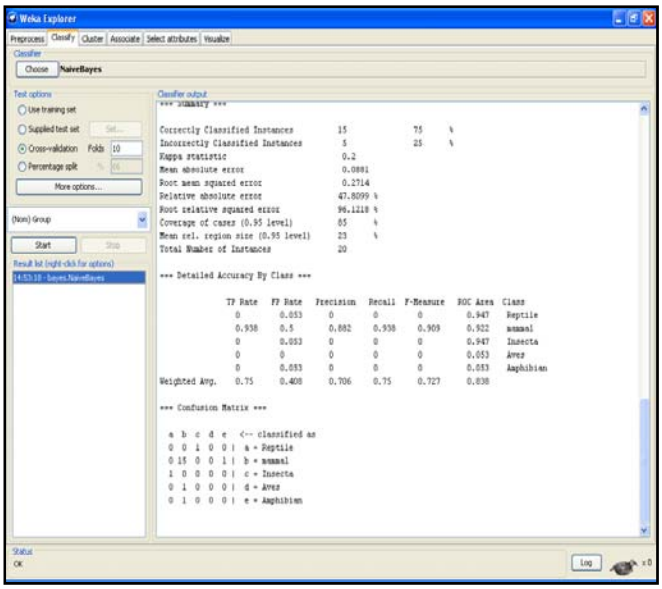

FIG. 2 SIMULATION RESULT FOR TESTING SET: NAÏVE BAYES

SVM:

Support Vector Machine classifier separates a set of objects into their respective groups with a line [14]. Hyper plane classifiers separate objects of different classes by drawing separating lines among the objects. Support Vector Machine (SVM) performs classification tasks by constructing hyper planes in a multidimensional space [11]. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. Training in SVM always finds a unique global minimum [13].

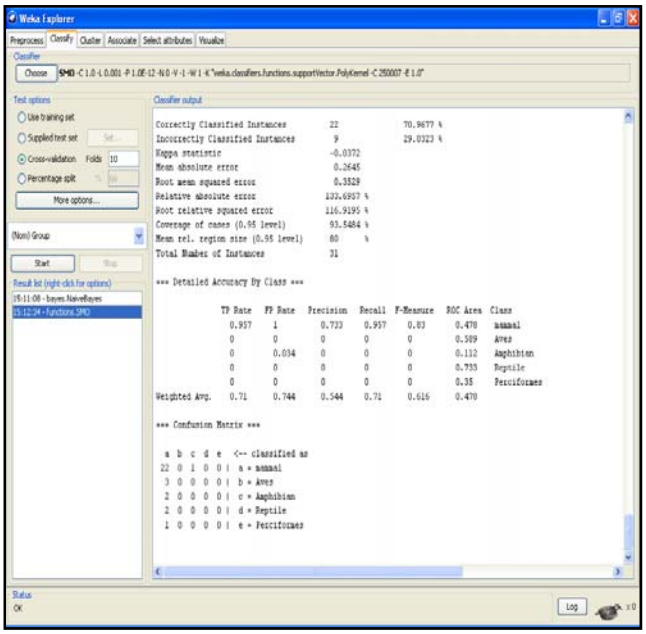

FIG. 3 SIMULATION RESULT FOR TRAINING SET: SVM

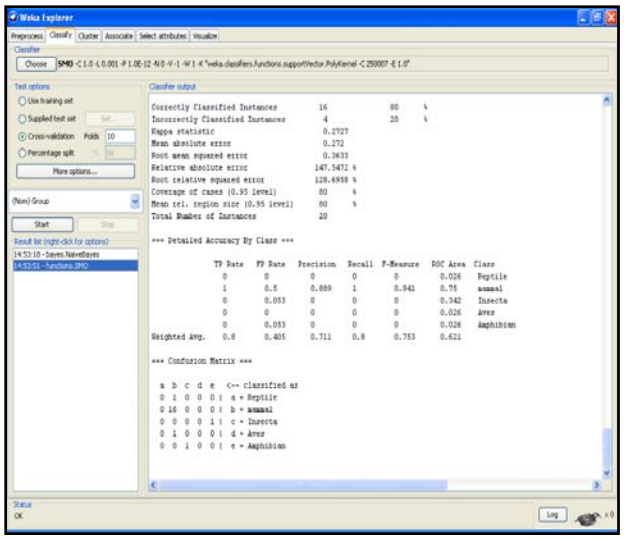

FIG. 4 SIMULATION RESULT FOR TESTING SET: SVM

IBK:

K-NN is a supervised learning algorithm, where a given data set is partitioned into a user specified number of clusters, K [9]. Predict the same class as the nearest instance in the training set. Training phase of the classifier stores the features and the class label of the training sets. New objects are classified based on the voting criteria [13]. It provides the maximum likelihood estimation of the class. Euclidean distance metrics is used for assigning objects to the most frequently labelled class. Distances are calculated from all training objects to test object using appropriate K value [15]. In this paper K value is assigned to 1 which shows that the chosen class label was the same as the one of the closest training object.

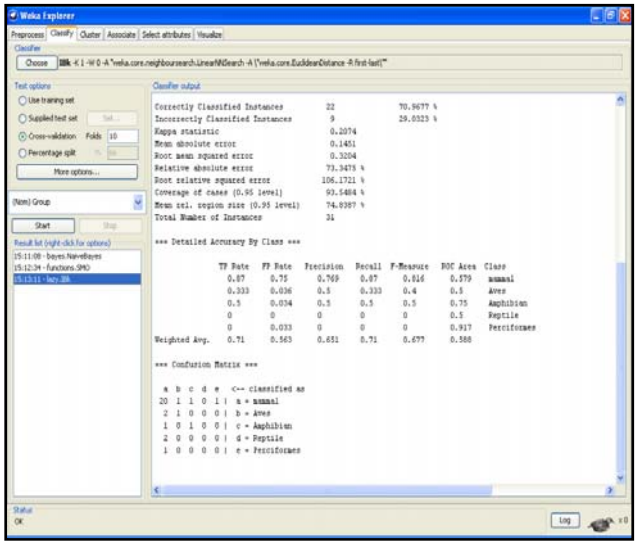

FIG. 5 SIMULATION RESULT FOR TRAINING SET: IBK

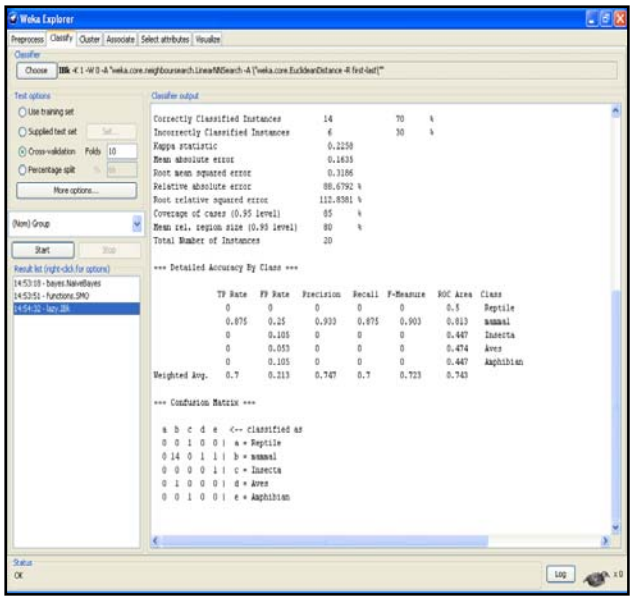

FIG. 6 SIMULATION RESULT FOR TESTING SET: IBK

J48 Classifier divides the training objects with a missing value. It provides fractional parts proportional to the frequencies of the observed non missing values [21]. Cross validation is used to split the data sets into training and testing. It builds decision trees from a set of training and testing data. At each node of the tree, classifier chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The attribute with the highest normalized information gain is chosen to make the decision. This algorithm then recurses on the smaller sub list of the data sets.

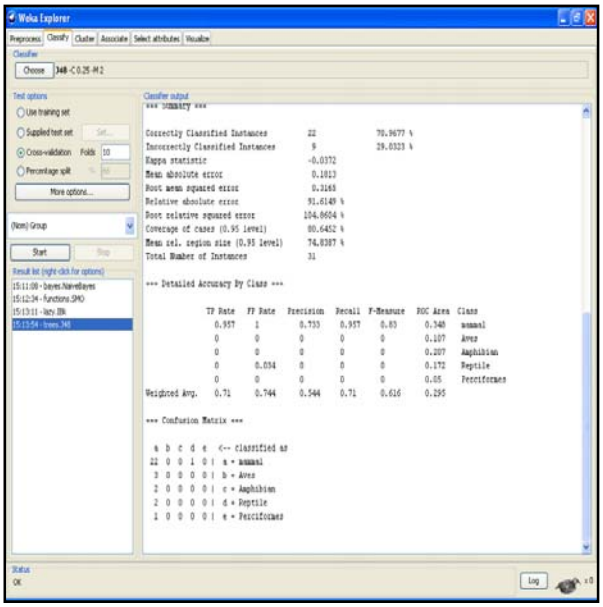

FIG. 7 SIMULATION RESULT FOR TRAINING SET: J48

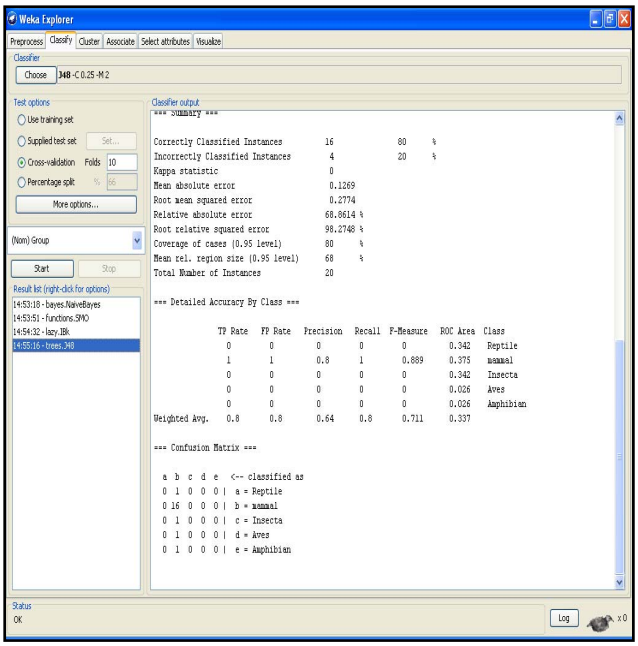

FIG. 8 SIMULATION RESULT FOR TEST SET: J48

Performance Evaluation

10-fold cross-validation technique is used to evaluate the performance of classification methods. Data set was randomly sub divided into ten equal sized partitions. Among the partitions nine of them were used as training set and the remaining one is used as a test set. Evaluation of performance is compared using Mean absolute error, root mean squared error and kappa statistics [18]. Large test sets gives a good assessment of the classifier's performance and small training sets which result in a poor classifier.

TABLE 1 CLASSIFIED INSTANCES FOR ANIMAL KINGDOM DATA SET

Performance rate / classifier	Correctly classified instances		Incorrectly classified instances	
	Training set %	Test set %	Training set %	Test set %
Naive Bayes	58.0645(18)	75(15)	41.355(13)	25(5)
SMO	70.9677(22)	80(16)	29.0323(9)	20(4)
IBK	70.9677(22)	70(14)	29.0323(9)	30(6)
J48	70.9677(22)	80(16)	29.0323(9)	20(4)

Kappa Statistics

Kappa is a normalized value of agreement for chance agreement.

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

Where P(A) = percentage agreement

P(E) = chance agreement.

If K =1 agreement is perfect between the classifier and ground truth.

If K=0 indicates there is a chance of agreement.

TABLE 2 KAPPA STATISTICS FOR TRAINING AND TEST SET FOR ANIMAL KINGDOM

Classifier	Kappa statistics	
	Training set %	Test set %
Naïve Bayes	-0.0372	0.2
SMO	-0.0372	0.2727
IBK	0.2074	0.2258
J48	-0.0372	0

Each classifier produces K value greater than 0 (i.e.) each classifier is doing better than chance for training set [5]. J48 classifier proves there is a chance of agreement. In the case of test set IBK classifier alone produce K value greater than 0, while other classifiers provide less than 0. Therefore compared to both training and test set j48 works better for training set and IBK works better for test set.

Mean Absolute Error

The mean absolute error (MAE) is a quantity used to measure predictions of the eventual outcomes. The mean absolute error is given by

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n} |e_i|$$

The mean absolute error is an average of the absolute errors $e_i = |f_i - y_i|$,

Where f_i = prediction

y_i= true value.

Root Mean Squared Error

Root mean squared error is the square root of the mean of the squares of the values. It squares the errors before they are averaged [18] and RMSE gives a relatively high weight to large errors.

The RMSE Ei of an individual program *i* is evaluated by the equation:

$$Ei = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\frac{P_{(ij)} - T_j}{T_j}\right)^2}$$

Where $P_{(ij)}$ = the value predicted by the individual program

i = fitness case

T_j =the target value for fitness case *j*.

TABLE 3 ERROR RATE FOR CLASSIFIED INSTANCES

Error rates / Classifier	Mean Absolute Error		Root Mean Squared Error	
	Training set %	Test set %	Training set %	Test set %
Naive Bayes	0.1718	0.0881	0.3609	0.2714
SMO	0.2645	0.272	0.3529	0.3633
IBK	0.1451	0.1635	0.3024	0.3186
J48	0.1813	0.1269	0.3165	0.2774

Training a data set generally mininmizes the error rate for test set. Error rate for training set is comparatively higher than that of the test set. From the above diagram IBK has the lowest error rate compared to other three algorithms. If both the algorithm has the same mean absolute error rate then root mean squared error rate is taken into consideration for choosing the best classification algorithm.

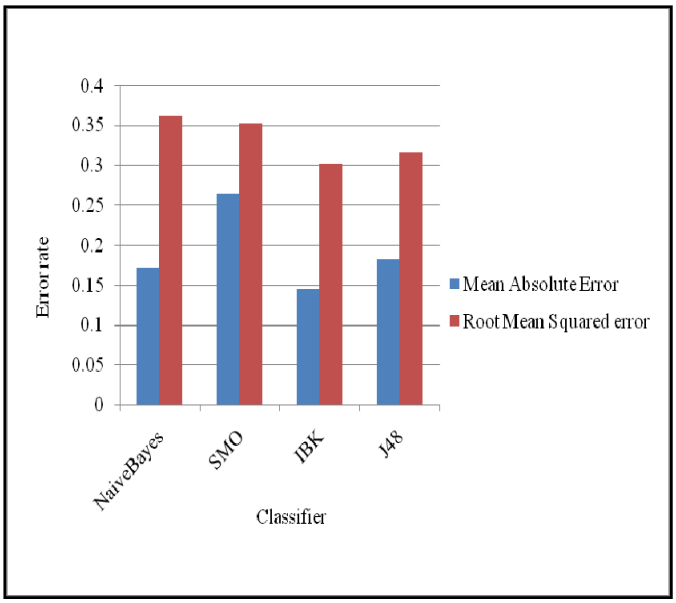

FIG. 9a ERROR RATE FOR TRAINING SET

Testing set has low error rate than the training data set. It is clear from the above diagram for the animal kingdom test set that Naive Bayes classifier has the lowest mean absolute error rate.

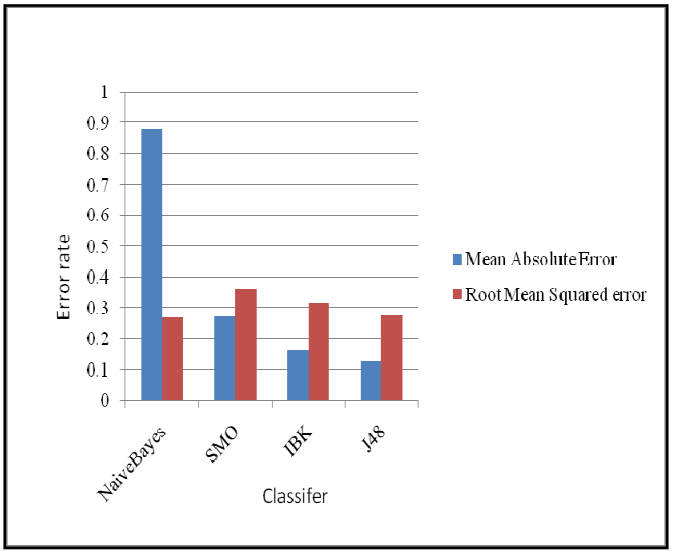

FIG. 9b ERROR RATE FOR TESTING SET

Confusion Matrix Classification Accuracy

Classification accuracy is the degree of correctness in classification. The degree of correctness is evaluated using various classifiers for individual instances in the animal kingdom data set. The Larger the training set and the higher the classifier accuracy is ; the smaller the test set and the lesser the classifier accuracy is Similarly larger test set provides a good assessment on classifier accuracy [17]. In this paper animal kingdom training set is higher than the test set which gives higher accuracy rate. Training set contains 60% of the whole data set and the remaining is used as test set for classification [21]. Remove Useless filter removes the unwanted attributes which reduces the time taken to build the model.

TABLE 4a: CLASSIFICATION ACCURACY RATE FOR CONFUSION MATRIX: TRAINING SET

Classifier / Animal Kingdom	Naive Bayes	SMO	IBK	J48
Mammal	22.5806	70.9677	32.2581	29.0323
Aves	87.0968	25.8065	90.3226	87.0968
Amphibian	87.0968	90.3226	96.7742	25.8065
Reptile	90.3226	96.7742	80.6452	90.3226
Perciforms	93.5484	67.7419	67.7419	9.6774

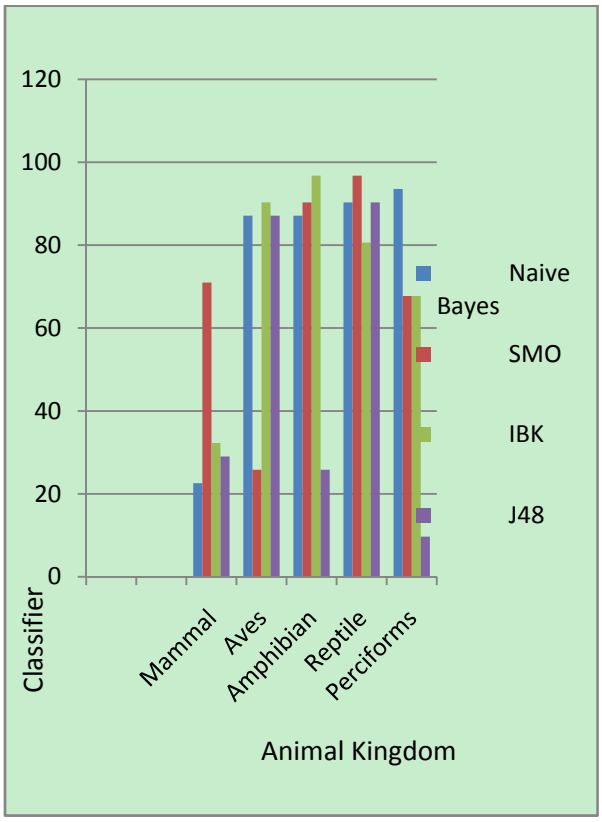

FIG. 10a ACCURACY RATE FOR TRAINING SET

SMO and IBK have the same accuracy rate performance compared to all other classifier algorithms [12]. This shows that the two algorithms are effective in classifying the training set. J48 provides the least result in classification. This classification accuracy rate depends upon the number of animal kingdom in the data set. For Mammal animal kingdom SMO has the highest accuracy rate for confusion matrix. IBK classifier has the highest accuracy rate for Aves and Amphibian animal kingdom. SMO has the highest accuracy rate for reptile animal kingdom. NaiveBayes shows the highest performance for perciforms [25].

TABLE 4b: CLASSIFICATION ACCURACY RATE FOR CONFUSION MATRIX: TEST SET

Classifier / Animal Kingdom	Naive Bayes	SMO	IBK	J48
Mammal	55	90	85	20
Aves	90	50	90	45
Amphibian	90	90	85	45
Reptile	90	75	5	65
Perciforms	90	90	85	65

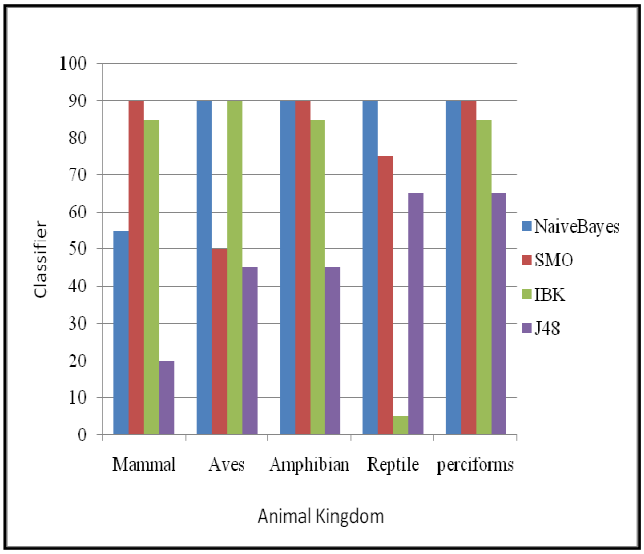

FIG. 10b ACCURACY RATE FOR TEST SET

Naive Bayes has higher accuracy rate for Aves, Amphibian, Reptiles and Perciforms animal kingdom in the above diagram. SMO has higher accuracy rate for Mammal, Amphibian and Perciforms. IBK has higher accuracy rate for Aves. J48 has considerable performance in Reptile and Perciforms animal kingdom data set.

Result and Discussion

The algorithm which has the lowest mean absolute error and higher accuracy is chosen as the best algorithm. If two algorithms show the same error rate and accuracy then the two algorithms are considered to be effective in classification. In this classification, each classifier shows different accuracy rate for different instances in the data set. SMO and IBK have the highest classification accuracy. Though both the same accuracy IBK the lowest mean absolute error compared to SMO. If both the algorithm have the same error rate and accuracy then root mean squared error is taken into consideration. SMO and IBK have the same correctly classified instances. 70.9677% for training set and 80% for testing set. Taking mean absolute error and classification accuracy IBK is considered as the best classification algorithm. Compared with training and test set J48 classifier is the least performing algorithm for the animal kingdom data set.

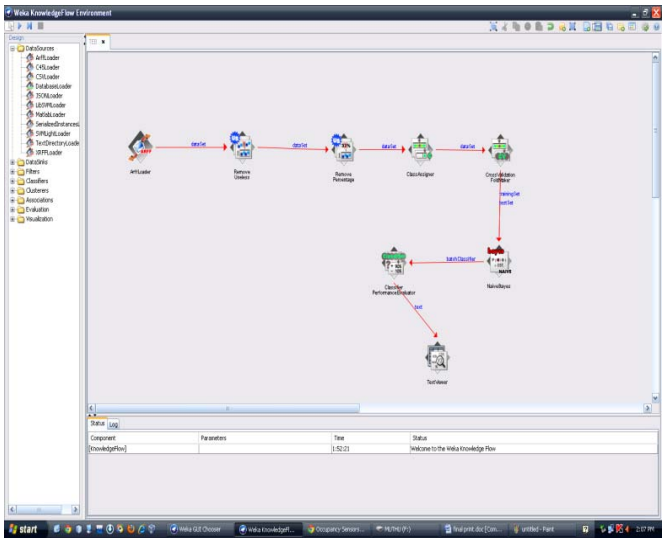

FIG. 11 KNOWLEDGE FLOW ENVIRONMENT DIAGRAM FOR ANIMAL KINGDOM DATA SET FOR NAIVE BAYES

Data flow diagram for the animal kingdom data set is verified using knowledge flow experimenter. The above figure shows the flow of the data set from the loader to the output. The output obtained from the explorer in weka is as such in experimenter and the output is verified.

Conclusion

This classification is discussed for evolutionary things in the environment. In this paper performances of the classifier are discussed for animal kingdom data set with respect to accuracy rate and mean absolute error and also Root Mean Squared Error. Training set and test set performance evaluation is also discussed. The best and worst classification algorithms are evaluated for training and test set. These best performing algorithms are used in case of evolutionary data set. For animal kingdom data set IBK is the best performing and J48 classifier is the least performing algorithm. This type of classification is applicable for population evolution, stock market changes data set, vehicle data set with various error measures.

REFERENCES

A.Blum, and P.Langley, "*Selection of Relevant Features and Examples in Machine Learning*," Artificial Intelligence, 1997.

A.Berson, S.Smith, K.Thearling, "Building *Data Mining Applications for CRM* ", in International Journal of Information Technology & Computer Science 2000.

A.Cufoglu, "*A Comparative Study of Selected Classifiers with Classification Accuracy in User Profiling*" in IEEE Conference 2009.

C.N.Silla, "*Novel top-down approaches for hierarchical classification and their application to automatic music genre classification*" in IEEE Conference 2009.

C.Sugandhi, P.Yasodha, M. Kannan."*Analysis of a Population of Cataract Patients Databases in Weka Tool*", in International Journal of Scientific and Engineering Research 2011.

David Meyer, "*Face Detection with Facial Features and Gender Classification Based On Support Vector Machine*", in International Journal of Imaging Science and Engineering 2010.

F.Gorunescu, "*Data Mining: Concepts, models and techniques*", Blue Publishing House, 2006.

G. Nguyen, Hoang, S. Phung, & A. Bouzerdoum, "*Efficient SVM training with reduced weighted samples*", in IEEE World Congress on Computational Intelligence 2010.

H.Jiawei, and K.Micheline, "*Data Mining-Concepts and Techniques*", in Elsevier Publishers, 2008.

H.M. Noaman, "*Naive Bayes Classifier based Arabic document categorization*" in IEEE Conference 2010.

J. R. Quinlan, "*Improved use of continuous attributes in c4.5*", in Journal of Artificial Intelligence Research, 1996.

Kappa at http://www.dmi.columbia.edu/homepages/chuangi/kappa.

K.Forster, "*Incremental KNN Classifier Exploiting Correct-Error Teacher for Activity Recognition*" in IEEE Conference 2010.

Li Rong, "*Diagnosis of Breast Tumor Using SVM-KNN Classifier*" in IEEE Conference 2010.

M.Dash, and H.Liu, "*Feature Selection for Classification*", Intelligent Data Analysis, 1997.

M.Govindarajan, RM.Chandrasekaran, "*Evaluation of K-nearest neighbor classifier performance for direct marketing*", in Expert system with applications 2010.

M.Julie, David and Kannan Balakrishnan, "*Significance of Classification Techniques in Prediction of Learning Disabilities*", in International Journal of Artificial Intelligence & Applications 2010.

N.Gayathri, A.V.Reddy and latha, "*Performance Analysis of Data mining Algorithms for Software Quality Prediction*" in IEEE Conference 2009.

Ravikumar.B, "*Comparison of Multiclass SVM Classification Methods to Use in a Supportive System for Distance Relay Coordination*" in IEEE Transaction 2010.

R.Kohavi, and G.H.John, "*Wrappers for Feature Subset Selection*," Artificial Intelligence, 1997.

Shengtong Zhong "*Local-Global-Learning of Naive Bayesian Classifier*" in IEEE conference 2009.

S.B.Kotsiantis, "*Supervised Machine Learning: A Review of Classification Techniques*", in IEEE Transaction 2007.

S.Belciug, "*Bayesian classification vs. k-nearest neighbour classification for the non-invasive hepatic cancer detection*", in International conference on Artificial Intelligence and Digital Communications 2008.

T.Darrell, and P.Indyk and G. Shakhnarovich, "*Nearest Neighbor Methods in Learning and Vision: Theory and Practice*". MIT Press 2006.

WEKA at http://www.cs.waikato.ac.nz/~ml/weka.